行政院國家科學委員會專題研究計畫成果報告

無母數參數估計在 IRT 真分數等化法的應用與分析

Nonparametric ICC Estimation with Application to IRT True Score Equating

計畫編號：NSC 89-2118-M-018 -005
執行期限：89 年 8 月 1 日至 90 年 7 月 31 日

主持人：李信宏　彰化師範大學數學系
E'mail: li@math.ncue.edu.tw
計畫參與人員：王雅苓　彰化師範大學數學系

一、中文摘要

等化是指利用統計方法來轉換兩個或者多個試卷分數的過程。這裡所指的多個試卷，實際上其測驗內容以及所欲測量的潛在能力是一致的，可以視為同一測驗的數個版本（same test but different forms）。這些試卷可以在不同日期施測，或者在同一時間對不同受試者施測，藉由等化過程，這些試卷的測驗結果亦即受試者的表現，就可以直接互相比較。

一般常用的等化方法包括了傳統的線性等化、百分位數等化以及 IRT 真分數等化和觀察分數等化等。其中 IRT 真分數等化乃是經由累加的試題反應模式函數值，來求取兩個試卷的同等分數，在這個轉換分數的過程中，必須使用統計方法來估計試題反應模式中的試題參數。一般而言，參數估計大部分採用聯合最大概似估計法、邊際最大概似估計法或者貝氏估計法等。這些方法一方面要考慮到試題反應模式的選擇是否適用於測驗資料，另一方面由於概似函數的複雜，試題參數估計值大都是藉由數值方法遞回計算來獲得，其結果應用於等化可能無法轉換成比較精確的同等分數。

本研究建議使用kernel smoothing來估計試題特徵曲線並應用於IRT真分數等化。此方法是一種無參數的分析方法，不需要先假設某一模式，而是根據實際作答結果，選擇適當的權數將作答資料做加權平均，進而估計試題特徵曲線。此外，計畫中也將設計各種不同狀況來實施模擬研究，特別是當測驗資料並不符合二參數以及三參數對數模式時，以強調使用無母數方法的優點。最後，研究的成果將和使用BILOG 3軟體估計參數進而實行等化的結果對照分析，希望藉此充分瞭解計畫所提方法的各種性質，進一步而能夠應用於實際的測驗資料。

關鍵詞：等化、IRT 真分數等化、試題反應模式、kernel smoothing、BILOG 3

Abstract

Equating is a statistical process that is used in situation where several alternate forms of a test exist and scores earned on different forms are compared to each other. Currently, a number of procedures have been developed in equating. For example, linear equating, equipercentile equating, IRT true score equating as well as observed score equating. Typically, after the item parameters are estimated and converted to be on the same scale, the IRT true score equating can be employed to transform scores between different forms. However, this method is based on strong model assumptions, which likely do not hold precisely in certain real testing situations. The proposed procedure used nonparametric regression methods such as kernel smoothing to estimate item

characteristics curves (ICCs) for each item. The estimated ICCs then are added together resulting in an estimated true score from which the IRT true score equating can be utilized. A full-scale simulation study is design to investigate the performance of our procedure, especially for equating errors. Meanwhile, the comparisons of behaviors of our method with other equating approaches is reported and carefully discussed.

Keywords: Equating, Item Response Theory (IRT) Model, IRT True Score Equating, Kernel Smoothing

二、計畫緣由與目的

自九十學年度起教育部將廢止高中聯招，全面實施多元入學方案，以基本學力測驗作為學生申請、參加推薦甄試或分發進入高中(職)或五專的參考資訊之一，以取代現行高中聯考制度。基本學力測驗的試題是採取事前命題，並經過多次修題、審題、預試與試題分析，以得到試題的相關訊息，然後將符合要求的試題納入題庫中。未來正式施測時，則會依據事前公佈的測驗目標，從題庫中抽取試題組成正式測驗進行考試。基本學力測驗和現階段各級聯考制度最大不同之處在於它可以一年多次實施，而參加不同次測驗的考生，其測驗分數可利用等化(equating)的方法，將不同次的測驗分數換算成同一個量尺的分數，使各次的測驗分數能放在同一個標準上來進行比較，由此例可看出等化的意義與應用。

IRT真分數等化(true score equating)的方法乃是建立於試題作答模式(item response model)。所謂的試題作答模式是將受測者能力與試題作答結果間的關係以數學模式表示，如果將此模式所要表達的關係以圖形表示時，則稱為試題特徵曲線(item characteristic curve , ICC)。在進行IRT真分數等化前，要先估計所選擇之試題作答模式的參數，通常使用LOGIST (Wingersky , Barton & Lord,1982)或BILOG (Mislevy & Bock,1986)等電腦軟體去估計每一題之試題參數及考生之能力參數。不過在運用這些軟體時必須先假設試題作答模式，例如一、二和三參數對數模式等，但實際上，不見得測驗中每一試題之作答結果皆符合上述既定模式，所以估計出來之試題特徵曲線就有可能不會接近真正之曲線，如依據不適當之模式進行等化時，等化結果就可能不是很理想。因此，本研究建議使用kernelsmoothing來估計試題特徵曲線。kernelsmoothing方法是選擇適當的函數作為權數(weights)，再根據實際作答資料做加權平均以估計ICC，此方法並不需要假設適當的模式，乃是一種無參數(nonparametric)分析方法。如此根據實際作答資料估計所得之ICC比較能符合真正之試題特徵曲線，再利用此較適當之ICC進行IRT真分數等化，並與選擇對數參數模式(3PL)做IRT真分數等化的結果做比較。

三、結果與討論

在模擬研究中，主要是比較在IRT真分數等化時，使用不同方法估計試題作答模式對等化結果的影響。我們比較以kernel smoothing估計試題特徵曲線和選擇3PL對數參數模式兩種情況，以循環方式將Form X上的分數等化至Form Y的分數，然後將Form Y的分數等化至Form Z的分數，再由Form Z等化至Form X (Gafni & Melamed, 1990)，最後結果是將Form X等化至Form X，所以在Form X之原始分數與在Form X之等化分數的誤差值(bias)愈小愈好。研究中以圖形來比較兩種情形之等化結果，並計算試卷上所有整數分數等化後的誤差平均值(mean bias)及均方差(mean square error，MSE)。另外，並考慮測驗長度(25題和40題)、受測者人數(750人、1500人和3000人)與猜測參數對等化結果的影響。

在進行主要模擬研究前，並先討論平滑參數h的選擇對kernel smoothing估計試題特徵曲線的影響。依據ACT Math test且令c=0.15模擬一份40題的試題參數，再藉由這些試題參數模擬3000位受試者的作答資

料。令h分別為0.1, 0.2, 0.3, 0.4, 0.5, 0.7，估計在不同h下之試題特徵曲線，並計算它們的誤差平方期望值的平方根(RMSE)。

由模擬研究的結果中可看出人數對smoothing等化結果的影響，人數愈多時，smoothing和3PL等化結果的結果都愈好。這個結果十分符合預期，因為人數愈多時kernel smoothing和BILOG估計的試題特徵曲線愈準確，所以等化情形也較好。至於測驗長度對等化的結果的影響，以模擬25題與40題的情況比較起來，兩種情形等化結果蠻相近的，25題的情況略為好些，但不是很明顯，所以在本研究中不能很明確的判斷測驗長度是否會影響等化結果，可能是這兩種題數相差不大的緣故。另外在h的選擇部份，利用RMSE選擇的結果，發現在人數較多時，所選擇的h的值愈小。例如，在750人和1500人時選擇h=0.3；而在3000人時h=0.2。但在猜測參數不存在的情況下，選h=0.2。

再進一步分析發現：運用smoothing的方法做真分數等化時，關於h的選擇是利用RMSE來做比較，選擇使大多數估計試題之RMSE最小的h_0，但選擇的此h值是對大多數試題而言好，對少數試題來說，此h值並非最理想，若是對每一試題皆採用不同的h，也許可以更增加所估計之試題特徵曲線的正確性。另外，等化部分在低分的結果較不理想，研究中採行的改進做法是去掉部分考生的作答結果再做等化，但如此是否會失掉一些資訊，是值得再討論的；而在高分部份有些結果也不是很理想，但是並沒有找出較好之改進辦法，若仿照低分部份的處理，則無法找出高分部份對應之能力值以進行等化，所以並沒有實施修正，這一部份可以作為日後研究的課題。

四、計劃成果自評

測驗等化是以統計方法建立同一測驗之不同試卷間的等化分數，使得在不同試卷上所得的分數也可以互相比較。而自九十年度起開始實施國民中學學生基本學力測驗，於不同次考試所得的分數須藉由測驗等化的方法將其轉換到同一量尺上以進行比較，因此測驗等化的工作也就愈顯重要。

應用IRT於等化時，試題作答模式的選擇是一重要的關鍵，如何選擇適當的試題作答模式，使其能正確描述受測者能力與試題作答間的關係是最為重要的。即使是常用的三參數對數模式，在實際情況中也不見得適用於每一測驗試題。因此我們建議使用kernel smoothing的方法來估計試題特徵曲線，也就是不先預設其作答模式，而依據其作答結果，選擇適當的權數將其結果做加權平均，以期能最真實的描述出試題特徵曲線，並完全呈現考生作答資料的資訊。經由模擬研究發現，選擇以kernel smoothing估計所得之試題特徵曲線作為等化的模式，大部份比以參數對數模式作為等化模式來的好。所以利用kernel smoothing的方法估計ICC，可以避免選擇不適當的模式，進而運用於IRT真分數等化中，更能增加等化結果的正確性。

五、參考文獻

[1] Gafni, N., & Melamed, E. (1990). Using the circular equating paradigm for comparison of linear equating models. *Applied Psychological Measurement,* 14, 243-256.

[2] Kolen, M.J., & Brennan, R.L. (1995). *Test Equating : Methods and Practices* New York : Springer.

[3] Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG : Item analysis and test scoring with binary logistic models.* Mooresville, IN : Scientific Software Inc.

[4] Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic *curve estimation . Psychometrika,* 56, 611-630.

[5] Wand, M.P., & Jones, M.C. (1995). *Kernel smoothing*. London : Chapman & Hall.

[6] Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton, N. J. : Educational Testing Service. Test equating (pp. 9-49). New York: Academic.